Preprints of the 7th IFAC Symposium on
Fault Detection, Supervision and Safety of Technical Processes
Barcelona, Spain, June 30 - July 3, 2009

Prediction of Vacuum Pump Degradation in Semiconductor Processing

Shane W. Butler * **John V. Ringwood** * **Niall MacGearailt** **

* *Department of Electronic Engineering, National University of Ireland, Maynooth, Co. Kildare, Ireland (e-mail: shane.butler@eeng.nuim.ie, john.ringwood@eeng.nuim.ie).*
** *Intel Ireland, Leixlip,Co. Kildare, Ireland (e-mail: niall.macgearailt@intel.com).*

Abstract: This paper addresses the issue of vacuum pump degradation in semiconductor manufacturing. The ability to identify the current level of vacuum pump degradation and predict the Remaining-Useful-Life (RUL) of a dry vacuum pump would allow manufacturers to schedule pump swaps at convenient times, and reduce the instances of unexpected pump failures, which can incur significant costs. In this paper, artificial neural networks are used to model the current level of pump degradation using pump process data as inputs, and a double-exponential smoothing prediction method is employed to estimate the RUL of the pump. We also demonstrate the benefit of incorporating process data, from the upstream processing chamber, in the development of a solution.

Keywords: Fault detection, neural networks, process models

1. INTRODUCTION

Almost all semiconductor manufacturing processes require some level of vacuum to operate. Ultra-high vacuum in a processing chamber is generally achieved using a turbo-molecular pump located at the processing tool followed downstream by a dry vacuum pump usually located within the subfab environment. Dry pumps are generally very reliable, but when applied to the pumping of particularly harsh processes in semiconductor manufacturing, they can occasionally suffer from unexpected failures.

Growing pressure on profit margins has lead IC manufacturers to increase their focus on improving process yields, tool uptime, and wafer throughput. At the same time, the increase in the use of tools for 300mm wafers and the introduction of new material technologies and cleaning flows constantly raise the by-product challenge and increase the value of wafers (Mooney and Shelley, 2005). The occurrence of a vacuum pump failure can cause irreparable damage to wafers, but also results in significant tool downtime and cleanup, which can can be a major expense. Furthermore, a vacuum pump failure results in unplanned maintenance of a pump which is significantly more expensive than planned maintenance in terms of resources, planning and manpower. As semiconductor manufacturing becomes an increasingly lean operation, vacuum pump suppliers are likewise expanding their operations in the field of condition monitoring and predictive maintenance.

Dry vacuum pumps typically comprise two separate pumps each driven by their own motor. The standard configuration consists of a single stage booster pump backed by a multi-stage main pump. These types of pumps are subject to a number of potential failure modes. These include sudden ingestion of deposits, exhaust pressure blockages, deposition causing pump seizure and the degradation of pump components leading to excessive loss of performance.

Several publications have addressed the issue of condition monitoring and predictive maintenance of vacuum pumps. (Mooney and Shelley, 2005) provides an overview of new capabilities in pump predictive maintenance through the introduction of networked monitoring systems. The issue of process by-products accumulating in the pumping mechanism was considered by (Konishi and Yamasawa, 1999). The accumulation of deposits within the running clearances of the pump causes friction resulting in the pump current exceeding current limits and causing the pump to shut down. In (Konishi and Yamasawa, 1999), the use of an ARMAX model to predict vacuum pump motor current was considered. The use of fuzzy-logic based condition monitoring is considered in (Twiddle et al., 2008), where a fuzzy-model based diagnostic scheme to detect mechanical inefficiency and exhaust system blockage in a dry vacuum pump is designed. It is demonstrated that the power ratios of certain frequency components in the exhaust pressure signal spectrum can be used to predict the gas load, motor current, and hence, mechanical efficiency.

The application of condition monitoring techniques to vacuum pump maintenance represents a significant technical challenge. Dry vacuum pumps are designed to have high reliability, very low maintenance, and the capability of pumping corrosive and reactive gas mixtures. However, modern semiconductor fabrication facilities operate multiple processes with such different operating conditions as

⋆ This work was supported by Enterprise Ireland under grant EI/CTFD/05/IT/323.

varying chamber pressures, gas flow rates, and different gas mixtures and properties. These process-related properties and operating conditions are often proprietary and as such, they are often inaccessible to the vacuum pump suppliers, (Cheung et al., 2006).

In the case study presented here, the degradation of a dry vacuum pump leading to excessive loss of vacuum performance is addressed. Pump data from a major semiconductor manufacturing facility is employed to develop a means to identify, track and predict the rate of pump degradation. The ability to identify and predict the loss of vacuum performance will allow for the planning of 'Just-In-Time' (JIT) pump exchanges at convenient times, resulting in less tool downtime and loss of product.

2. PROBLEM DESCRIPTION

During its lifetime, a dry vacuum pump is exposed to large quantities of often toxic and corrosive gases. In some cases, the gases used may reduce the expected lifetime of the pump. Semiconductor processing chambers require cleaning between each processed wafer. A common gas used for this task is nitrogen trifluoride (NF_3). This gas is highly corrosive, and in large quantities, can lead to an increased rate of degradation of dry pump components, resulting in the gradual loss of vacuum performance. Eventually, this loss of performance results in the foreline pressure in the processing chamber exceeding tolerance limits and the possibility of irreparable damage to wafers. This paper focuses on a means to identify the current level of pump degradation from analysis of pump variables, and having identified the current level of degradation, provide a means to estimate the Remaining-Useful-Life (RUL) of the pump.

The proposed solution comprises two elements. A diagnostic element to determine the current level of pump degradation and a prognostic element to provide an estimate of the RUL of the pump, given the current degradation level and rate of degradation. Artificial Neural Networks (ANNs) are employed to model the level of pump degradation, and a Double Exponential Smoothing Prediction (DESP) method is used to estimate the RUL of the pump.

3. DATA COLLECTION

3.1 Pump Data

In this study, dry pump data from 14 processing chambers in a major semiconductor manufacturing facility was available. Each of the chambers run a similar deposition process. Pump data covering approximately one full year of operation is available. The recorded data includes variables such as current, power, temperature and exhaust pressure.

Each of the chambers employ an Edwards (formerly BOC Edwards) iH600 dry vacuum pump. The iH600 pump comprises an HCDP80 dry pump, with an HCMB600 mechanical booster pump fitted to the inlet of the HCDP80 pump. Both the HCDP and HCMB pumps have enclosed, water-cooled, motors.

Pump data collection was carried out using the Edwards FabWorks system. FabWorks is a condition monitoring and analysis system for vacuum and exhaust management equipment. The FabWorks system is capable of monitoring up to 3000 items of equipment connected across its network. Each piece of equipment sends both status and process variable update information to a central server where the information is stored and processed. Access to the FabWorks system can be made from any computer connected to the LAN, allowing for both real-time and historical analysis of pump data.

The design of and installation of each individual network is dependant on the number of pieces of equipment connected to it. This is undertaken to ensure sufficient bandwidth is always maintained for the transmission of data across it. One of the design features of such a network is the selection of the data sampling approach. In this case, the variables in each pump are recorded using an event-based sampling technique known as send-on-delta sampling or, delta-logging (Mooney and Shelley, 2005). Using this approach, each variable is sampled using a regular sampling interval. Once the latest variable is sampled $x(t_i)$, it is compared to the most recent value sent to the network database $x(t_{i-1})$. If the difference between the two values exceeds or equals a preset threshold value δ, as in equation 1, then the latest variable value is timestamped and sent to the central server. If the threshold value δ is not exceeded, then no update is sent to the server.

$$|x(t_i) - x(t_{i-1})| \leq \delta \tag{1}$$

This sampling approach is very suitable for monitoring of dry pumps in a large semiconductor fabrication facility, which may have in excess of 1000 dry pumps in operation. A consideration in the installation of such a network is the rate at which pump data and status updates are sent across the network so that sufficient bandwidth is always maintained.

A major factor influencing the rate at which variable updates occur is determined by the selection of the δ value for each variable. This value effectively acts as a trade-off between signal tracking accuracy and network bandwidth. The smaller the δ value, the greater the signal tracking accuracy and network load, and vice versa.

3.2 Chamber Data

In addition to the pump data available for the study, process data from the upstream chambers in the form of status information and foreline pressure measurements was made available. The status information provided a means to determine exactly the times at which the pump is processing gases from upstream and when it is idling (processing no gases). In some cases, it is possible to determine when the pump is processing gases from analysis of the pump variables, the load on the pump increases when processing gases. However, depending on the load generated and the signal tracking resolution in the pump data, this may not always be possible, or reliable. Knowledge of when the pump is processing provides a means to identify the pump data, idling, and processing modes.

The foreline pressure measurements allow for a means to determine the current level of the degradation in the pump and to quantify the loss of performance. Figure 1 plots the average foreline pressure for each processed wafer observed during the final seven months of a dry pumps lifetime. Figure 1 demonstrates how, as the pump degrades over time, the average foreline pressure in the chamber rises. Eventually, acceptable tolerance limits for foreline pressure in the processing chamber are exceeded and processing ceases.

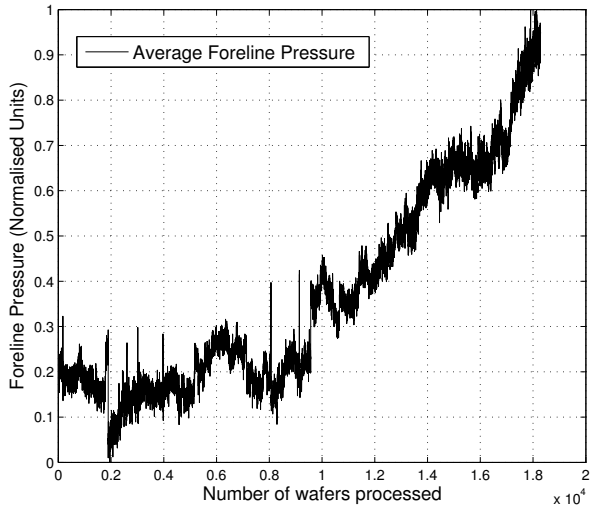

Fig. 1. Average foreline pressure for each processed wafer

During wafer processing, foreline pressure measurements are recorded at a much faster sampling rate than the pump data, and generate vast quantities of data. A means to extract the most useful data needs to be considered. Semiconductor wafer processing is generally undertaken via a number of 'recipe steps' using different gases, flow rates and pressure settings. In this application, the average foreline pressure over a single identified recipe step was employed to represent the average foreline pressure during the processing of a single wafer. The specific recipe step was selected for a number of reasons.

- the recipe time step was of significant duration relative to the total wafer processing time
- the chamber gate valve remained completely open and was not used to control chamber pressure during the recipe step

Hence, the major factor influencing foreline pressure during the recipe step should be the performance of the dry pump.

Due to the proprietary nature of the process chamber data, each of the data streams for foreline pressure from each chamber are re-scaled automatically so that it is not possible to determine a physical value from the data. To overcome this issue, the foreline pressure measurements from each chamber over the lifetime of the pump were mapped to the interval [0,1]. In this study, the rescaled foreline pressure values represent the level of degradation in the pump.

In general, vacuum pump suppliers do not have access to such upstream performance data, and certainly not in real-time, and as such, a major driver for this study is to identify any benefits of incorporating such upstream process data in the development of algorithms for vacuum pump condition monitoring. By having access to and incorporating such data in the development phase, it will be demonstrated how pump degradation can be identified and predicted using only the available pump data in real-time.

4. DATA PREPROCESSING

4.1 Data Reconstruction

The use of an event-based sampling approach to dry pump data collection results in the data stream for each pump variable being on an irregular sampling interval. The time between the receipt of updates of variable values can be significant, depending on the variable in question, the pump state and the selection of the δ value.

To overcome this issue, the pump data was reconstructed onto a regular sampling interval. A one minute sampling interval was chosen, for two reasons. No updates of temperature values were ever identified occurring, within one minute of the most recent update received, thus ensuring all temperature updates received were incorporated on the one minute interval. The temperature signals were reconstructed using linear interpolation. The power signals were first resampled onto a two second interval, which was sufficient to incorporate 98% of updates received. The power signals were reconstructed such that, if no variable update was received within the two second interval, the most recently available value was used. The power signals were then averaged over each minute to correspond with the temperature values. In addition, a one minute interval was chosen as the wafer processing time was an integer multiple of one minute, at approximately eight minutes. This allowed the pump data to be easily tabulated with the chamber foreline pressure and status data.

Having reconstructed the data onto a regular sampling interval, statistical summary data was produced in the form of mean values for each observed variable, during the processing of each wafer. The use of mean values per wafer has the advantage of reducing the signal noise introduced by the reconstruction of the data onto a regular interval. Furthermore, it allows for the processing of the data on a wafer per wafer basis, and for any RUL prediction to be made in terms of number of wafers until failure.

4.2 Data Filtering

Due to the data reconstruction approach, significant signal noise was introduced to the pump process data. The generation of mean values for each variable on a per wafer basis served to reduce the signal noise somewhat. However further filtering to smooth the data was necessary. An Exponentially Weighted Moving Average (EWMA) filter

was employed. The form of the filtering is shown in equation (2) below.

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1} \quad (2)$$

The EWMA filter applies a weighting factor which decrease exponentially. The weighting for older data points decreases exponentially, giving more importance to more recent observations, whilst not discarding older observations entirely. The degree of smoothing is determined by the selection of the smoothing constant α.

Figure 2 plots the normalised mean values per wafer of the drypump temperature of a pump approaching failure. Both the original data and the filtered data using the EWMA filter with $\alpha = 0.01$ are shown. The output of the filter serves to smooth the data significantly whilst still retaining all of the important trends in the data. Each of the variables were filtered this way.

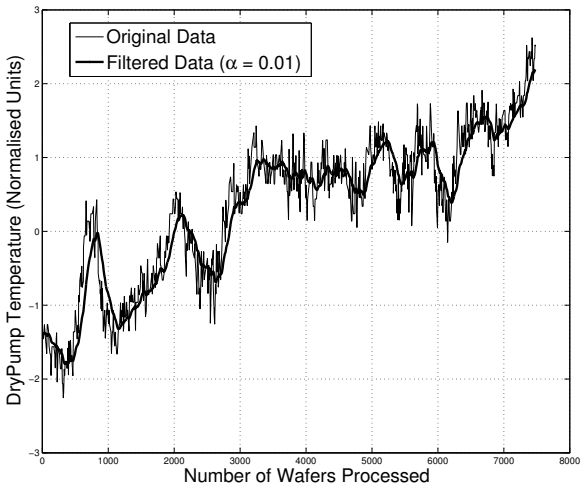

Fig. 2. Filtering applied to dry pump temperature data

5. MODELING PUMP DEGRADATION

The proposed solution to the issue of vacuum pump degradation comprises two stages. The first is a diagnostic element to determine the current level of pump degradation from analysis of process data from the pump. The second prognostic stage attempts to predict the current pump degradation level into the future in order to determine an estimate of the RUL of the pump.

The diagnostic element of the proposed solution uses an ANN to determine the current level of pump degradation using the summary process data from the pump as inputs to the network. The target vector in training for the network is the level of pump degradation identified from the foreline pressure data.

5.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a non-linear mapping techniques, inspired by biological neural networks. Just as in a biological system, ANN training involves adjustments to the synaptic connections (weights) between neurons.

An ANN can be trained for many applications such as function approximation (nonlinear regression), pattern association, or pattern classification. During training, the weights and biases of the network are iteratively adjusted using a training algorithm such as backpropagation to minimize the network performance function. In function approximation, for example, ANNs are trained, so that particular input values lead to specific target outputs. The network is adjusted, based on a comparison of the output and the target, until the network output matches the target.

ANNs have been applied to numerous problems in the fault diagnosis and system modeling domain (Venkatasubramanian et al., 2003). These include (Rietman, 1998), where ANNs are employed to model and predict changes in pressure in a plasma reactor, using process data as inputs.

In this study, a standard feed-forward MLP neural network was employed to model pump degradation. Various network architectures were considered and tested. The selected architecture comprises a single hidden layer with twenty neurons, using tan-sigmoid activation functions and a single output layer using a log-sigmoid activation function to limit the output to between [0,1]. The training data was split into training, validation and test data, to help prevent overfitting of the data. The number of hidden layers and neurons were chosen by testing of various combinations and the network was trained using the Levenberg-Marquardt backpropagation algorithm.

An input feature vector was generated for training of the ANN. The vector comprises averaged values of temperature and power for both the booster and the main pump, over the period of a single processed wafer. The selection of these four variables to form the input vector, was based upon the correlation of their filtered values with the foreline pressure. The inclusion of the current signal for the booster and main pumps resulted in a significant deterioration in performance and so were not included in the final input feature vector.

The rescaled mean foreline pressure values per processed wafer, were used as the target values for the training of the network. The objective of the network was to identify the current level of pump degradation on a per-wafer basis, by analysis of the summary pump data. In this study, complete historical data for four pump failures as a result of vacuum pump degradation was available. Network training was carried out using combinations of three of the datasets as inputs, using the remaining dataset for testing.

5.2 Network Testing and Results

Figure 3 compares the modeled degradation level against the actual degradation level using a dataset from a failed pump. The dataset was not used during the network training process. In this case, the model output tracks the

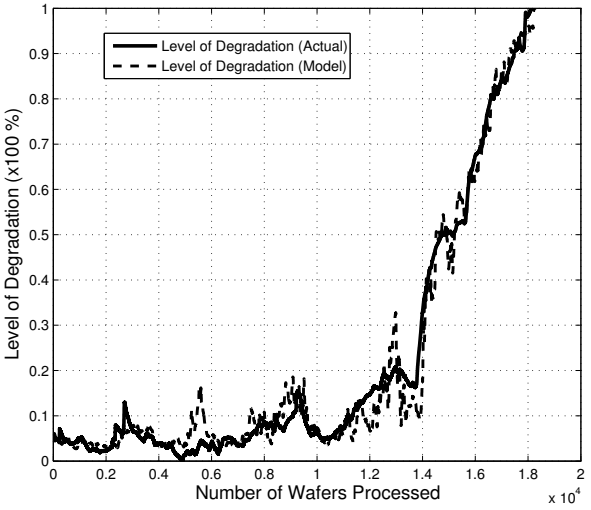

Fig. 3. Modeled and actual level of pump degradation

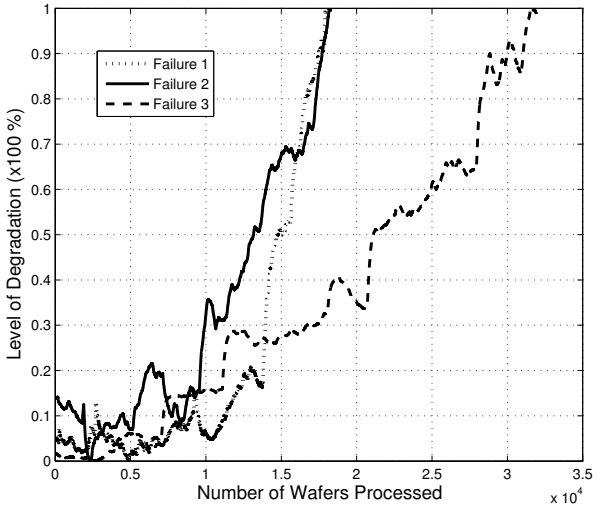

Fig. 4. Degradation trends for observed pump failures

actual level of degradation quite accurately, demonstrating that the current level of pump degradation can be identified from analysis of the pump data.

6. RUL ESTIMATES

Having demonstrated the ability to determine the current level of pump degradation, the most useful information to the manufacturer is an estimate of the RUL of the pump. This allows for planning of pump replacements at a convenient time, with minimal disruption to manufacturing. In determining the RUL of the pump, there are a number of factors which cannot be determined *a priori*, such as the future utilisation rate of the pump (idling/processing) and the number of times the pump is shutdown and restarted, which can influence the RUL.

Figure 4 shows degradation trends for three pump failures. Analysis of the available data, suggests that the failure trend generally becomes well established by the time the pump has approached 80% degradation. The times to failure observed between degradation exceeding 80% and pump failure ranged from approximately 1000 to 3500 wafers. This represents between 150 and 500 hours, approximately, of wafer processing. Predictions of RUL at and above 80% level of degradation were considered appropriate, to provide a sufficient horizon for taking corrective action.

As the level of pump degradation approaches 80%, the time series observations of the pump degradation levels are 'generally' of the form in Equation 3. Double exponential smoothing-based prediction models a given time series using a simple linear regression equation, where the y-intercept β_0 and the slope β_1 are slowly changing over time. In such cases, double exponential smoothing can used to apply unequal weighting to the individual elements of the time series.

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{3}$$

In order to obtain updated estimates of the time series, double exponential smoothing uses what is called the single and double smoothed statistic, S_T and $S_T^{[2]}$. These values are computed using two smoothing equations, (4) and (5), where both equations use the same smoothing constant α, which lies within the range [0,1]. This value determine the degree of smoothing applied to the data. The first equation smoothes the original time series and the second filters the S_T values.

$$S_T = \alpha y_T + (1-\alpha) S_{T-1} \tag{4}$$

$$S_T^{[2]} = \alpha S_T + (1-\alpha) S_{T-1}^{[2]} \tag{5}$$

Using S_T and $S_T^{[2]}$, updated estimates of β_0 and β_1 are determined as:

$$\beta_0(T) = 2S_T - S_T^{[2]} - T\beta_1(T) \tag{6}$$

$$\beta_1(T) = (\frac{\alpha}{1-\alpha})(S_T - S_T^{[2]}) \tag{7}$$

Having estimated β_0 and β_1, the forecast made at time T for the future value of $y_{T+\tau}$ is given by,

$$\hat{y}_{T+\tau}(T) = \beta_0(T) + \beta_1(T)(T+\tau) \tag{8}$$

With some manipulation, see (Bowerman and O'Connell, 1993), it can be shown that the forecast $\hat{y}_{T+\tau}(T)$ is given by,

$$\hat{y}_{T+\tau}(T) = (2 + \frac{\alpha\tau}{(1-\alpha)})S_T - (1 + \frac{\alpha\tau}{(1-\alpha)})S_T^{[2]} \tag{9}$$

Such an approach to estimating the RUL of a degrading system has previously been applied to determining time-to-wash intervals for shipboard gas turbine engines which experience gradual performance degradation caused by the ingestion of salt (Kacprzynski et al., 2001).

The choice of a number of parameters must be carefully considered in the application of the double exponential smoothing prediction. The smoothing constant α is determined by simulated forecasting of an historical dataset. Using a section of available historical data, a regression line

is fitted to the data and the initial least-squares estimates of β_0 and β_1 are determined. Using (6) and (7), initial values of S_T and $S_T^{[2]}$ can be calculated. Using these values, the historical dataset is used to update S_T and $S_T^{[2]}$ and in each time period a forecast is computed using the current values of S_T and $S_T^{[2]}$. The procedure is repeated for a range of α values, and the value which minimises the sum of the squared forecast errors in selected for use in forecasting future values.

The time series generated by the output of the neural network generally comprises thousands of values, where the difference between each consecutive value is quite small. In order to extract the relevant linear trend from the data over a smaller dataset, a new dataset comprising every 50th value of the time series was complied. This dataset was used for the prediction of the estimated RUL of the pump, and the generation of approximate 95% confidence limits.

Figure 5 illustrates the performance of the DESP method applied to the prediction of the RUL of a pump from 80% observed degradation. Also shown is the actual rate of degradation in the pump, and the approximate 95% confidence limits of the prediction.

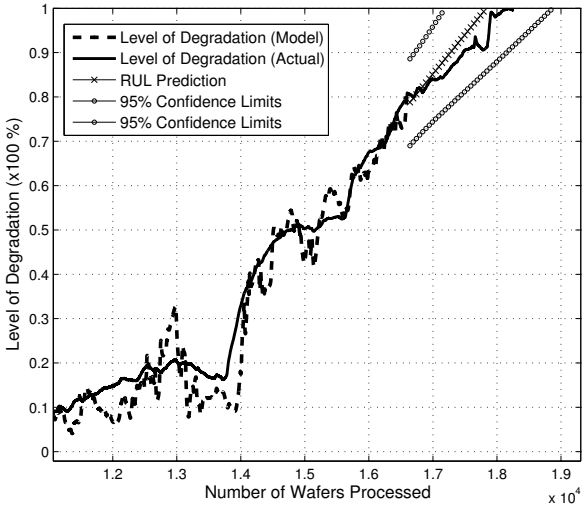

Fig. 5. RUL Prediction at 80% observed degradation

As can be seen, the actual time of failure of the pump fell well within the confidence limits for the prediction. The DESP method predicted pump failure at approximately 1250 wafers into the future with confidence intervals in the range of 590 to 2300 wafers. The actual time of pump failure was 1700 wafers into future.

7. CONCLUSIONS

In this paper, we have presented a method to both, identify the current level of vacuum pump degradation, and to estimate the RUL of a dry vacuum pump. The developed solution has the potential to reduce the instances of unexpected pump failures caused by pump degradation. This could help reduce pump maintenance costs, but also, the costs associated with scrapped wafers, tool downtime and chamber cleaning, following an unexpected pump failure.

In general, vacuum suppliers do not have access to process data from the upstream processing chambers. In this paper, we have demonstrated how, by incorporating status data and foreline pressure measurements from the upstream processing chamber in the development of a solution, a means to quantify the level of pump degradation was possible. In the absence of such data, a means to determine accurately the current level of pump degradation on those manufacturing tools is now possible.

As the database of historical pump failures grow, a number of issues may be considered. A reduction in the δ values may greatly improve the signal resolution and present opportunities for improving the model accuracy and RUL prediction. A larger database may also present an opportunity for considering alternative approaches to modeling the level of pump degradation.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of Edwards (formerly BOC Edwards) in carrying out this work.

REFERENCES

Bruce L. Bowerman and Richard T. O'Connell. *Forecasting and Time Series: An Applied Approach*. Duxbury Press, 1993.

W. Cheung, J. Lim, K. Chung, and S. Lee. A precision diagnostic method for the failure protection and predictive maintenance of a vacuum pump, June 2006. URL `www.wipo.int/pctdb/en/wo.jsp?wo=20060649911`.

G. Kacprzynski, M. Gumina, M. Roemer, D. Caguiat, T. Galie, and J. McGroarty. A prognostic modeling approach for predicting recurring maintenance for shipboard propulsion systems. In *ASME Turbo Expo*, New Orleans, USA, Jun 2001.

S. Konishi and K. Yamasawa. Diagnostic system to determine the in-service life of dry vacuum pumps. *IEE Proceedings - Science, Measurement and Technology*, 146:270–276, 1999.

M. Mooney and G. Shelley. Data collection and networking capabilities enable pump predictive diagnostics. *Solid State Technology*, 48:49–63, 2005.

Edward A. Rietman. A study on failure prediction in a plasma reactor. *IEEE Transactions on Semiconductor Manufacturing*, 11, 1998.

J. A. Twiddle, N. B. Jones, and S. K. Spurgeon. Fuzzy model-based condition monitoring of a dry vacuum pump via time and frequency analysis of the exhaust pressure signal. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 222:287–293, 2008.

V. Venkatasubramanian, R. Rengaswany, S. N. Kavuri, and K. Yin. A review of process fault detection and diagnosis, part III: Process history based methods. *Computers and Chemical Engineering*, 27:327–346, 2003.